LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-225442

The BlueGene/L Supercomputer and Quantum ChromoDynamics

P. Vranas, R. Soltz

October 20, 2006

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

UCRL-TR-225442

The BlueGene/L Supercomputer and Quantum ChromoDynamics

Update for the Gordon Bell competition entry gb110s2

Dear Gordon Bell Competition Committee,

Since the submission of our paper we have made some very important and significant improvements. As well, we have also included in our competition entry a new member. Our colleague Ron Soltz from LLNL has joined us in the effort to obtain the best scaling and sustained speed for QCD. He has been instrumental in this effort. Please add him and his affiliation (Ron Soltz, Lawrence Livermore National Laboratory, soltz@llnl.gov) as a member of our Gordon Bell competition entry.

In summary our update contains:
1) Perfect speedup sustaining 19.3% of peak for the Wilson D-slash Dirac operator.

2) Measurements of the full Conjugate Gradient (CG) inverter that inverts the Dirac operator. The CG inverter contains two global sums over the entire machine. Nevertheless, our measurements retain perfect speedup scaling demonstrating the robustness of our methods.

3) We ran on the largest BG/L system, the LLNL 64 rack BG/L supercomputer, and obtained a sustained speed of 59.1 TFlops. Furthermore, the speedup scaling of the Dirac operator and of the CG inverter are perfect all the way up to the full size of the machine, 131,072 cores (please see Figure II). The local lattice is rather small (4x4x4x16) while the total lattice has been a lattice QCD vision for thermodynamic studies (a total of 128x128x256x32 lattice sites). This speed is about five times larger compared to the speed we quoted in our submission.

As we have pointed out in our paper QCD is notoriously sensitive to network and memory latencies, has a relatively high communication to computation ratio which can not be overlapped in BGL in virtual node mode, and as an application is in a class of its own. The above results are thrilling to us and a 30 year long dream for lattice QCD.

In more detail (corresponding to the above summary):
1) The code produces intermediate data (the spin projected spinors) that need to be stored in memory but are not reused until later. As a result they do not need to be stored into the L1 cache. In fact it is better if they are not stored in L1 because they will inadvertently cause L1 evictions and as a result tax the memory bandwidth. Therefore, we store this data via a separate TLB that sets the L1 cache to write-through, store without allocate mode. The system software team provided us with a function that gives a memory "window" with these L1 attributes. This function was used to produce the performance numbers in our paper but it has not been available in the system software present in the large BG/L installations until very recently. This function has now become part of the standard IBM BG/L system software Release 3. Release 3 was recently installed in the IBM Watson BG/L 20 rack supercomputer. Similarly we use a function that allows us access to the small on-chip SRAM. We use this fast memory to perform the on-chip core-to-core communications by local copy. Again this function is now available in Release 3.

Using the above and a local lattice with four-dimensional size 4x4x4x16 we produced the speedup graph of the Wilson D-slash Dirac operator in figure I below (blue). The largest sustained speed is 22.25 TFlops with perfect scaling and sustained performance of 19.3% of peak.

2) We measured the speedup of the full Conjugate Gradient (CG) inverter that inverts the Dirac operator. As you know there are two global sums over the full machine per CG iteration. This causes a small drop of sustained performance from 19.3% to 18.7% . However, as you can see in figure I (magenta), the speedup scaling is still perfect indicating the robustness of our methods.

3) We ran our code on the LLNL 64 rack BG/L system. Since Release 3 of the BG/L system software is very recent it is not yet available in that system. As a result our sustained performance for the Dirac operator is 16.1% and for the full CG inverter is 15.9%. As you can tell from figure II the speedup is perfect all the way up to the largest machine size of 131,072 cores. The largest sustained speed for the Dirac operator is 59.1 TFlops and for its CG inverter 58.0 TFlops. This is unprecedented speed and scalability for an exceptionally difficult and important application.

The enormous size of this machine (32x32x64x2 cores) results in a lattice of 128x128x256x32 sites. This size lattice is of extreme interest to the physics of QCD thermodynamics and is a "dream" lattice for lattice gauge theory. Also, please note that our local lattice size of 4x4x4x16 is rather small and is exposed to all communications and memory latencies as described in the paper. There are no large sequential accesses here. A typical communication/memory access chunk is between 100 to 1000 Bytes.

Please consider the 59.1 sustained TFlops as our largest speed for the Dirac operator and 58.0 sustained TFlops for the CG inverter. Also, please consider graphs I and II below as replacing figure 2 in our original submission.

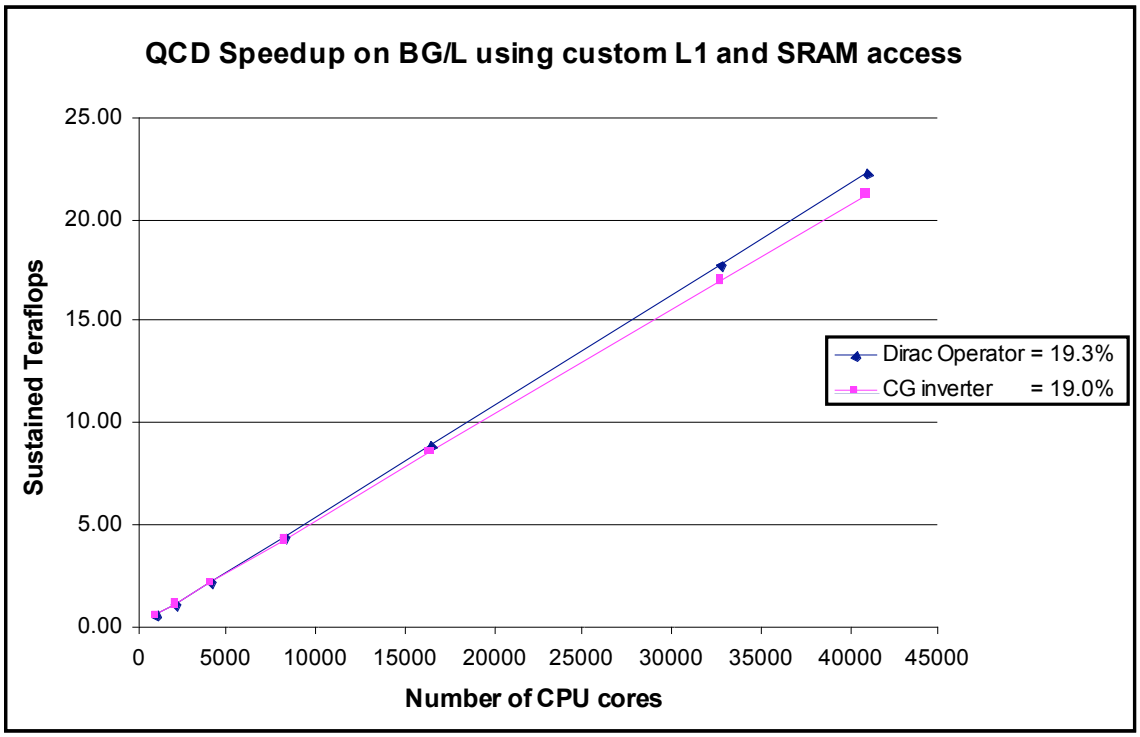

Figure I

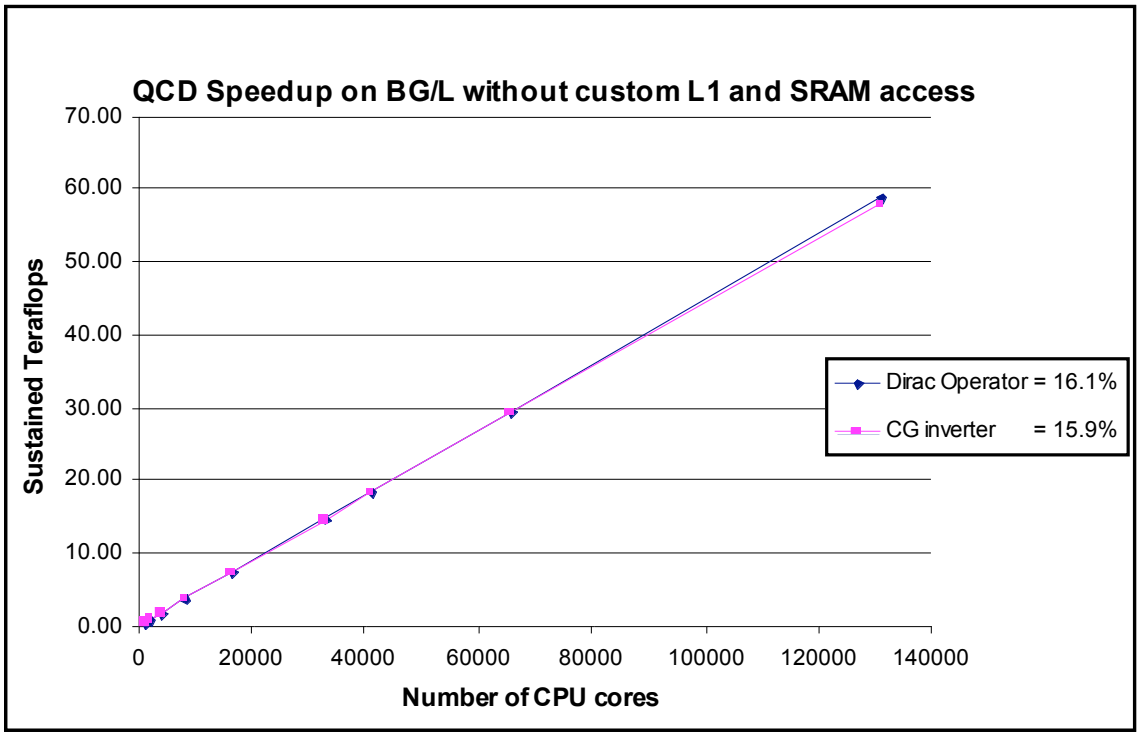

Figure II

Machines:
All performance numbers up-to 20 racks (40,960 cores) were obtained in the IBM Watson 20 rack BG/L supercomputer. The 32 racks (65,536 cores) and 64 racks (131,072 cores) numbers were obtained in the LLNL 64 rack BG/L supercomputer.

Acknowledgements:
We are grateful to IBM Watson and to Lawrence Livermore National Laboratory for allowing us access to these most precious resources.